第 2 章要点——复杂 HMTL 解析

1. HTML 和 （以 http://www.pythonscraping.com/pages/warandpeace.html 为例）

- HTML 元素是块级元素，它是可用于组合其他 HTML 元素的容器，没有特定的含义，可用于对大的内容块设置样式属性。如： …
- HTML 元素是内联元素，可用作文本的容器，它也没有特定的含义，可用于为部分文本设置样式属性。如：

2．BeautifulSoup 的 findAll 和 find 函数可以获取指定标签

- findAll(tagName,tagAttributes, recursive,text,limit,keywords)
- find(tagName,tagAttributes, recursive,text,keywords)

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen("http://www.pythonscraping.com/pages/warandpeace.html")
bsObj=BeautifulSoup(html.read(),"html.parser")
nameList=bsObj.findAll("span",{"class":"green"})          or findAll(text=”the prince”)
for name in nameList:
    print(name)                                           or print(name.get_text())
```

3．BeautifulSoup 库里的四种对象

- BeautifulSoup 对象
- Tag（一个或一组标签）对象
- NavigableString（标签里的文字）
- Comment（HTML 文档的注释标签<!-- -->）对象

```
Name4=bsObj.findAll("span",{"class":"green"}) [4]
print(name4)
```

4．HTML 标签结构（以 http://www.pythonscraping.com/pages/page3.html 为例）

- 以表中第 2 行的图标签为例：Html-body-div-table-tr-td-img
- 子标签（children）是父标签的下一级
- 后代标签（descendants）是父标签下面所有级别的标签
- 兄弟标签（siblings）是同级的所有标签
- 父标签（parent）是上一级标签

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen("http://www.pythonscraping.com/pages/page3.html")
bsObj=BeautifulSoup(html,"html.parser")
```

```
for child in bsObj.find("table",{"id":"giftList"}).children:                    or descendants
    print(child)

for sibling in bsObj.find("table",{"id":"giftList"}).tr.next_siblings:          or previous_siblings
    print(sibling)

print(bsObj.find("img",{"src":"../img/gifts/img1.jpg"}).parent.previous_sibling.get_text())
```

5．正则表达式（regular expression）

- 返回符合表达式规则的字符串
- 字母 a 至少出现 1 次：a+；字母 b 重复 5 次：bbbbb；字母 c 重复任意偶数次：(cc)*
- 表 2-1，P25-P26（英文版），P22（中文版）
- 邮箱地址：[A-Za-z0-9\._+]+@[A-Za-z]+\.(com|org|edu|net)

6．正则表达式和 BeautifulSoup

- re 库，compile 函数
- 抓取页面上所有指定路径的图形标签
- 获取标签属性：myTag.attrs，返回的是字典对象

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import re
html=urlopen("http://www.pythonscraping.com/pages/page3.html")
bsObj=BeautifulSoup(html,"html.parser")
images=bsObj.findAll("img",{"src":re.compile("\.\./img/gifts/img.\.jpg")})
for image in images:
    print(image.attrs["src"])
```

7．Lamba 表达式

- lambda 表达式本质上是一个函数，只是没有函数名
- findAll 函数的参数可以是函数，这个函数须以标签为参数，返回布尔型

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html=urlopen("http://www.pythonscraping.com/pages/page3.html")
bsObj=BeautifulSoup(html,"html.parser")
tags=bsObj.findAll(lambda tag: len(tag.attrs)==2)
for tag in tags:
    print(tag)
    print()
```